구문 패턴과 키워드 집합을 이용한 통계적 자동 문서 분류의 성능 향상

한 정 기† · 박 민 규†† · 조 광 제††† · 김 준 태††††

요 약

본 논문에서는 통계적인 분류 방법과 지식 기반 분류 방법을 통합하여 정확도를 향상시키는 자동문서 분류 모델을 제시한다. 본 논문에서 제안하는 모델에서는 지식 기반 방법을 먼저 적용하고, 지식 기반 방법에 의해 분류되지 않은 문서에 통계적 방법을 적용한다. 이러한 통합된 방법을 사용함으로써 분류 실패 없이 모든 문서를 분류하면서 정확도를 향상시킬 수 있다. 통계적인 방법으로는 벡터 모델에 역카테고리빈도 가중치를 정의하여 사용하였고, 지식 기반 방법으로는 문장 패턴을 표현하는 방법으로 구문 패턴과 키워드 집합을 정의하여 패턴 매칭을 수행하였다. 일간지 기사를 대상으로한 실험을 통하여 두 방법을 통합하여 사용함으로서 분류 정확도를 높일 수 있음을 보였다.

Improving the Performance of Statistical Automatic Text Categorization by using Phrasal Patterns and Keyword Sets

Jung-Gi Han† · Min-Gyu Park†† · Kwang-Je Cho††† · Juntae Kim††††

ABSTRACT

This paper presents an automatic text categorization model that improves the accuracy by combining statistical and knowledge-based categorization methods. In our model we apply knowledge-based method first, and then apply statistical method on the texts which are not categorized by knowledge-based method. By using this combined method, we can improve the accuracy of categorization while categorize all the texts without failure. For statistical categorization, the vector model with Inverted Category Frequency (ICF) weighting is used. For knowledge-based categorization, Phrasal Patterns and Keyword Sets are introduced to represent sentence patterns, and then pattern matching is performed. Experimental results on news articles show that the accuracy of categorization can be improved by combining the two different categorization methods.

1. 서 론

문서 분류란 복수의 분류 카테고리를 정해놓고 문서의 내용에 따라 하나 또는 그 이상의 카테고리를 문서에 지정함으로써 문서들을 집단화하는 작업이다. 이러한 문서 분류는 대부분 수작업에 의해 이루어져 왔으나, 온라인 문서의 양이 점차 많아지고 그 종류가 다양해지면서 문서 분류의 자동화에 대한 필요성이 널리 인식되어지고 있다[1, 3, 4-6].

일반적으로 문서를 자동 분류하는 방법에는 문서 내에 나타나는 단어의 빈도를 이용하여 분류 카테고리를 찾는 통계적인 분류 방법(statistical categorization)[7, 8, 13, 14, 17, 18]과, 인간 전문가가 행하는 것처럼 문서의

※ 본 논문은 한국과학재단 핵심전문연구(과제번호 961-0904-024-2)의 연구비 지원에 의한 것임
† 정 회 원 동국대학교 대학원 컴퓨터공학과
†† 정 회 원 웹 패턴 테크놀로지 개발실 근무
††† 정 회 원 : 서울시스템 DTP사업본부 개발팀 근무
†††† 정 회 원 동국대학교 컴퓨터공학과 교수
논문접수 1999년 8월 18일, 심사완료 2000년 3월 3일

내용을 기반으로 하는 분류 규칙에 따라 분류를 수행하는 지식 기반 분류 방법(knowledge-base categorization)[7-9, 11]이 있다.

통계적인 분류 방법은 문장의 내용에 대한 분석 없이 단어의 출현 빈도만을 이용하여 문서를 분류함으로서 분류 방법이 간단하고 수행속도가 빠르지만 문서의 내용에 대한 분석을 수행하지 않으므로 분류의 정확도에 한계가 있다[8, 9] 지식 기반 방법은 문서의 내용을 기반으로 분류 규칙을 만들어 사용하므로 분류된 문서들의 경우 높은 정확도를 나타내지만 충분한 규칙을 제공하지 못하면 분류되지 못하는 문서들의 비율이 높을 수 있다는 단점이 있다.

본 논문에서는 자동문서 분류의 정확도 향상을 위하여 통계적인 분류 방법에 지식 기반 분류 방법을 복합적으로 사용하는 분류 방법을 제안한다. 통계적인 분류 방법으로는 기존의 벡터 유사도(vector similarity)에 의한 분류 방식에 분류 정확도를 향상시키기 위하여 역카테고리빈도(Inverted Category Frequency, ICF) 가중치를 정의하여 사용하였으며, 지식 기반 분류 방법으로는 구문 패턴(Phrasal Pattern)과 키워드 집합(Keyword Set)을 정의하여 간단한 자연언어처리를 수행하였고, 이 두 가지 방법을 복합적으로 사용하는 방안을 제시하였다.

본 논문에서 제시한 분류 방법들의 성능을 알아보기 위하여 조선일보 경제 기사를 대상으로 자동 분류 실험을 수행하였으며, 실험을 통하여 본 논문에서 제시한 통계적인 분류 및 지식 기반 분류 방법이 모두 효율적임을 보였고, 두 방법의 통합에 의하여 전체 분류 정확도를 향상시킬 수 있음을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 일반적인 문서의 자동 분류 방법에 대하여 설명하였고, 본 논문에서 사용한 통계적 분류 방법을 3장에, 문장 패턴을 이용한 지식 기반 분류 방법을 4장에, 그리고 두 가지 방법을 통합하는 방법을 5장에 설명하였다. 본 논문에서 제시한 분류 방법들에 대한 실험과 그 결과에 대하여 6장에서 설명하였으며, 결론 및 향후 과제를 7장에 제시하였다.

2. 문서의 자동 분류

문서의 자동 분류 방법을 크게 두 가지로 구분해 보면, 이미 분류되어 있는 문서들로부터 각 분류 카테고리에 나타나는 단어들의 출현 빈도에 대한 정보를 추출하여 분류에 이용하는 통계적인 방법과, 문서들이 가지고 있는 문장의 뜻을 파악하여 분류에 이용하는 지식 기반 방법이 있다.

통계적인 분류 방법은 사람에 의해 이미 분류되어 있는 문서들(training set)로부터 각 분류 카테고리에 나타나는 단어들의 출현 빈도에 대한 정보를 추출하고, 분류하고자 하는 문서로부터 주요 단어들과 단어들의 출현 빈도를 추출한 뒤 이러한 정보를 이용하여 가장 적합한 카테고리를 찾거나 각 카테고리에 대하여 포함 여부를 판단하는 것으로, 많이 사용되는 통계적인 분류 방법으로는 Bayesian Probability를 이용하여 문서가 각 카테고리에 속할 확률을 계산하는 방법[3, 14, 18]과, 분류하려는 문서와 각 카테고리에 포함된 문서들 간의 유사도를 계산하는 방법[5, 13, 17, 20]등이 있다.

Bayesian Probability를 이용한 분류 방법은 분류하려는 문서에 단어 $W_1, W_2, \ldots, W_n$ 나타난 경우, 각 단어가 나타나는 사건이 독립적이라고 가정하면, 이 문서가 카테고리 C_j 에 분류될 확률을 식(1)과 같이 계산한다.

$$P(C_j \mid W_1, W_2, \cdots, W_n) = k * P(C_j) * P(W_1 \mid C_j) * P(W_2 \mid C_j) \cdots * P(W_n \mid C_j) \quad (1)$$

이때 k는 모든 카테고리에 대한 계산에 공통으로 사용되는 비례상수이며, $p(C_j)$와 $p(W_i|C_j)$는 수작업에 의해 분류되어 있는 문서집단으로부터 근사치를 계산할 수 있다.

벡터 유사도 계산에 의한 분류 방법은 분류하려는 문서와 분류 대상 카테고리들을 단어들의 벡터로 나타내고, 두 벡터 사이의 유사한 정도를 비교하는 것으로 대표적인 유사도 계산 방법으로는 두 벡터 사이의 각도를 나타내는 코사인 계수가 있다 벡터 D가 분류하려는 문서를 나타내고, 벡터 C_j가 카테고리 C_j를 나타낸다고 하면 이 문서와 카테고리 C_j 사이의 유사도는 식(2)와 같이 계산한다.

$$Similarity(D, C_j) = \frac{D \cdot C_j}{|D||C_j|} \quad (2)$$

문서와 카테고리의 표현은 해당 문서 혹은 카테고리에 단어 W_i가 출현한 경우 i번째 원소를 1, 그렇지 않은 경우 0으로 하는 불리안 벡터로 나타낼 수도 있고,

강확도를 높이기 위해 각 단어에 상대빈도와 역문헌빈도(IDF)[2, 27]를 이용한 가중치를 부여할 수도 있다. 역문헌빈도는 적은 수의 문서에 나타나는 단어에 대해 높은 가중치를 주는 것으로, 상대빈도와 역문헌빈도를 이용한 경우 단어 W_i의 가중치는 단어 W_i의 문서(혹은 카테고리) j에서의 빈도수를 $freq_{ij}$, 문서의 개수를 N, 색인어 W_i를 포함하는 문서의 개수를 DF_i라 할 때 $freq_{ij}*(log(N)-log(DF_i)+1)$과 같이 계산된다

지식 기반 방법은 분류 대상 문서의 샘플들을 분석하여 분류 규칙들을 만들고 이러한 규칙을 이용하여 문서 분류를 수행하는 것으로, 문서의 내용에 따른 분류 규칙을 만드는 방법[6-9, 11]과, 문서 내용 외의 정보들을 이용하는 방법[1]이 있다.

문서의 내용에 따른 분류 방법으로는 특정 카테고리로의 분류에 결정적인 단서가 되는 핵심 단어들을 추출하여 이러한 단어들의 출현 여부에 따라 분류를 수행하도록 하는 방법, 그리고 특정 카테고리로 분류되는 문서들에 자주 나타나는 구나 문장 형태를 패턴으로 표현하여 패턴 매칭에 의해 문서를 분류하는 방법, 문서의 내용을 파악하여 문서를 분류하는 방법이 있다.

문서 내용 외의 정보들을 이용하는 방법으로는 문서의 작성 부서와 같은 정보들을 이용하는 규칙을 만들어 문서를 분류하는 방법으로 전문가 시스템 형태로 구현될 수 있다.

일반적으로 통계적인 방법은 단어들의 출현 빈도를 기반으로 각 카테고리로 분류될 확률이나 각 카테고리와의 유사도를 계산하므로 가장 높은 값을 갖는 단일 카테고리로 문서를 분류하는 경우, 모든 문서를 분류할 수 있으나, 문서의 내용을 분석하는 것은 아니므로 분류의 정확도에는 한계가 있다 지식 기반 방법은 사람에 의해 분류 대상 문서들에 대한 분석이 이루어 진 후 분류 규칙을 만들어 사용하므로 규칙에 따라 분류된 문서들의 경우 높은 정확도를 나타내지만 충분한 규칙을 제공하지 못하면 분류되지 못하는 문서들의 비율이 높을 수 있다는 단점이 있다.

따라서 이러한 상반된 특징을 가지는 두 가지 방법을 복합적으로 이용하면 두 방법이 상호 보완 작용을 하도록 하여 전체적인 분류 정확도를 향상시킬 수 있을 것이다. 본 논문에서는 벡터 유사도를 이용한 통계적인 분류 방법에 문장 패턴을 이용한 지식 기반 분류 방법을 접목하여 분류 대상 문서를 모두 분류하면서 분류의 정확도를 향상시키는 방법을 제안한다

3. 벡터 유사도에 의한 통계적인 분류

이 장에서는 본 논문에서 사용한 벡터 유사도에 의한 통계적인 분류 방법에 대하여 설명한다. 앞에서 설명한 통계적인 분류 방법들 중 IDF 가중치를 이용한 벡터 유사도 방법이 문서 분류에 많이 사용되고 있다. 그러나 일반적으로 정보 검색에서 사용하는 IDF 가중치를 문서 분류에 사용하는 경우 다음과 같은 문제가 있다. C_1, C_2를 분류 카테고리, D_1, D_2, D_3, D_4를 분류된 실험집단 문서, W_1, W_2를 색인어라 하고, 이들이 다음과 같이 분류되어 있다고 하자.

C_1 : D_1, D_3 $\quad$ W_1 : D_1, D_2에 나타남
C_2 : D_2, D_4 $\quad$ W_2 : D_1, D_3에 나타남

문서의 분류를 위해서는 W_2가 W_1보다 카테고리의 구분에 도움이 되므로 더 중요한 색인어라고 볼 수 있으나, IDF를 이용한 경우 두 단어의 가중치는 같게 되어 이와 같은 색인어의 특성을 반영하지 못한다. 본 논문에서는 통계적인 문서 분류를 위해 위의 예에서와 같이 카테고리의 분리 능력이 우수한 색인어에 높은 가중치를 주는 역카테고리빈도(Inverted Category Frequency, ICF)를 정의하고 이를 이용한 계층적 분류체계에서의 분류 방법을 제안한다

3.1 역카테고리빈도

본 논문에서 정의하는 ICF는 총 카테고리의 개수를 M, 색인어 W_i를 포함하는 카테고리의 개수를 CF_i 라고 할 때 식(3)과 같다

$$ICF_i = \log(M) - \log(CF_i) + 1 \qquad (3)$$

단어 W_i의 문서(혹은 카테고리) j에서의 빈도수를 $freq_{ij}$라고 하면 문서 j에서의 단어 W_i의 가중치는 $freq_{ij} * ICF_i$가 된다 분류체계가 평면적인 경우(단일 레벨로 구성된 경우)에는 이러한 방법으로 각 단어에 가중치를 준 다음 각 문서를 포함된 단어들의 가중치 벡터로 정규화 하여 표현하고, 각 카테고리를 각 카테고리에 이미 분류되어 있는 문서들(training set)의 평균 벡터로 나타내니 분류하려는 문서의 벡터 D와 각 카테고리 벡터 C_j 사이의 유사도를 2장에서 설명한 코사인 계수에 의해 계산하여 가장 높은 유사도를 갖

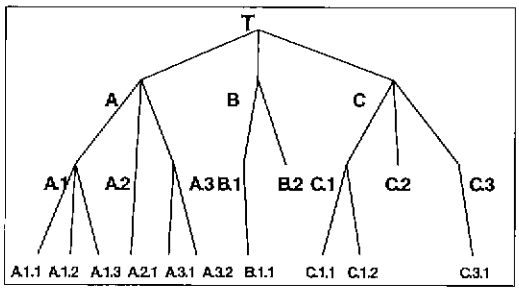
는 카테고리로 문서를 분류한다.

ICF는 IDF와 기본 원리는 같지만, IDF는 문서간의 분리도가 높은 단어에 높은 가중치를 주는 것이고, ICF는 카테고리간의 분리도가 높은 단어에 높은 가중치를 주는 것이라는 차이점이 있다. 즉 소수의 카테고리에 많이 나온 단어에 대해 높은 가중치를 주고, 여러 카테고리에서 고르게 나오는 단어에 대해서는 낮은 가중치를 주는 것이다.

문서 분류에 있어서는 카테고리간의 구분에 도움이 되는 색인어가 중요도가 높다고 할 수 있으므로, ICF가 IDF보다는 의미 있는 가중치 계산 방법이 된다. 특히 ICF는 카테고리들이 계층적 구조를 갖고 있을 경우, 어느 깊이까지 분류할 것인가를 결정하는데 있어서 현재까지 분류된 카테고리의 부속 카테고리(subcategory)들에 대해서만 ICF를 계산함으로서 분류의 정확도를 높일 수 있다.

3.2 계층적 분류

본 논문에서는 분류체계가 평면적이지 않고 계층적인 경우에 대하여도 문서의 분류 방법을 정의한다. 계층적 분류체계의 경우 평면적인 분류체계에서의 분류와 달리, 어느 정도 깊이까지 문서를 분류할 것인가 하는 문제가 발생한다. 다음과 같은 계층적 구조를 갖는 분류체계가 있다고 가정하자.



(그림 1) 계층적 카테고리

(그림 1)에서 만일 어떤 문서 D가 level 1(A, B, C로 구성)에서는 A, level 2에서는 $A.3$, level 3에서는 A.3.2와의 유사도가 가장 높다고 한다면 이 문서를 이 세 단계 중 어떤 깊이의 카테고리로 분류할 것인가를 결정해야 한다. 본 논문에서는 이와 같은 경우 임계값(threshold)을 사용하여 분류를 수행하는 다음과 같은 분류 방법을 제안한다. 최상위 레벨의 카테고리에서부터 부속카테고리들에 대해 유사도를 계산하여 가장 높은 결과를 갖는 부속카테고리를 선택하되 유사도가 임계값 이상일 때만 선택하도록 하며, 만약 부속카테고리들의 유사도가 임계값을 넘는 것이 하나도 없다면, 그 상태에서 분류를 멈추고 현재의 카테고리를 결과로 채택한다. 여기서 부속카테고리들에 대해 유사도 계산을 할 때는 현재 카테고리의 부속카테고리들만을 가지고 다시 ICF를 계산하게 된다, 즉 ICF 값이 항상 고정된 것이 아니라 레벨을 내려감에 따라 대상 그 부속카테고리들에 맞추어서 동적으로 계산되는 것이다. 위와 같은 방법을 알고리즘으로 기술해 보면 (그림 2)와 같다.

이 알고리즘에서 θ는 임계값을, T는 최상위 카테고리를 의미하며, $C_{result} = T$ 이면 분류에 실패한 경우를 뜻한다.

```
C_result = T, i = 1
WHILE (i ≤ max level)
    Compute ICF for all W_k against
            subcategories C_k of C_result
    Compute S_k = similarity(D, C_k) for
            subcategories C_k of C_result
    IF (all S_k < θ) RETURN (C_result)
    C_result = C_k with max S_k
    i = i + 1
```

(그림 2) 계층적 분류 알고리즘

4. 문장패턴을 이용한 지식 기반 분류

이 장에서는 본 논문에서 사용한 지식 기반 분류 방법에 대하여 설명한다. 앞서 언급한 바와 같이 통계적인 문서 분류는 문장의 내용에 대한 분석을 하지 않음으로 그 정확도에 한계가 있다. 이러한 문제점을 보완하기 위하여 문장의 의미를 명확히 나타내는 패턴들을 정의하여 이를 분류에 이용할 수 있다. 동일한 카테고리로 분류되는 문서들은 의미상 유사성을 가지므로 문서들 사이에 비슷한 형태를 갖는 문장들이 존재할 확률이 높다. 따라서 이러한 문장의 패턴들을 수집하고 이들을 새로운 문서에 매칭시켜 매칭된 패턴이 추출된 카테고리로 문서를 분류할 수 있다.

비슷한 의미를 가지는 문장이라고 할지라도 우리말과 같이 어순이 자유롭고 어미 변화가 다양한 언어에서는 그 형태가 여러 가지가 될 수 있다. 본 논문에서는 문장을 패턴화 시키는 방법으로 구문 패턴을 이용한 방법과 키워드 집합을 이용한 방법을 각각 정의하

여 사용하였다. 영어의 경우 제한된 영역의 문장으로부터의 정보추출에는 문장 패턴을 이용한 의미 분석 방법이 매우 효율적임이 이미 입증된바 있다[12].

4.1 구문 패턴

구문 패턴(Phrasal Pattern)에 의한 방법은 패턴 매칭에 사용할 문장 형태를 단어나 단어 집합들에 의한 정규표현(regular expression)으로 나타내고 유한상태 기계(finite state machine)에 의해 이들을 분류하려는 문서와 매칭하며 분류를 수행하는 것이다. 패턴의 정의에 사용되는 어휘들을 T_i라 하면 구문패턴에 의한 분류 규칙의 형태는 다음과 같다

$$\text{Phrasal Pattern Rule} : (T_1+T_2+\cdots+T_n) \rightarrow C_k$$

구문패턴을 정의하기 위하여 먼저 이미 분류되어 있는 실험집단의 문서들로부터 카테고리별로 분류의 단서가 되는 자주 쓰이는 표현을 찾는다. 예를 들어 경제 기사의 경우, '기업' 카테고리에 포함된 '부도'라는 소카테고리에서 자주 나타나는 표현들은 다음과 같다.

문장 1 : 요업개발은 경영악화와 자금난으로 회사가 파산에 직면했다.
문장 2 : 동창제지가 어음 20억원을 막지 못해 최종 부도 처리됐다.
문장 3 : 한국강관이 10일 서울 민사법원에 법정관리를 신청했다.

분류의 단서가 되는 구문들이 찾아지면 이들로부터 고유명사 등 분류의 단서가 된다고 볼 수 없는 요소들을 모두 제외시키고 핵심이 되는 명사와 동사들을 중심으로 '명사+조사', '동사+어미' 등이 연결된 형태로 패턴을 정의한다. 본 논문에서는 패턴을 정의하는데 사용하기 위해 조사와 어미를 의미에 따라 다음과 같이 분류하였다.

〈표 1〉 조사 및 어미의 분류

코드	조사 / 어미 그룹
J1	가, 이, 께서, 에서, 은, 는, 도
J2	를, 르, 을
J3	에, 에게, 한테, 께, 더러, 보고
J4	와, 과, 하고,
J5	로, 에게로, 으로
E1	하였다, 되었다, 중이다, 했었다
E2	었다, 었다고
E3	어, 고, 다, ㄴ다
E4	시키+ㄴ, 하+자, 하-르, 하+기, 하+ㄴ다, 되+ㄴ다, 되-ㄴ, 이+다

예를 들면, 앞에서 보인 문장 1에서 분류의 단서가 되는 부분은 '파산에 직면했다'라고 볼 수 있고, 해당되는 조사와 어미는 J3과 E1이므로 패턴은 (N^c J1 파산 J3 직면 E1)과 같이 정의할 수 있다. 패턴의 정의에 있어서 패턴의 복잡도와 사용할 패턴의 수는 분류의 정확도 및 분류율(패턴 매칭이 되어 분류되는 문장의 비율)과 밀접한 관계가 있다. 패턴이 단순해지고 수가 많아질수록 분류율은 높아지지만 정확도가 떨어지고, 반대의 경우에는 정확도는 높아지나 분류율이 낮아지게 된다.

문서의 분류는 문서내의 문장들을 각각 형태소 분석한 후 정의된 구문 패턴들과 패턴 매칭을 수행하여 매칭되는 패턴이 나타내는 카테고리로 문서를 분류한다. 정의된 패턴들을 이용하여 패턴 매칭을 수행했을 때 하나의 문서에 서로 다른 카테고리를 나타내는 패턴들이 매칭되는 경우에는 다수의 패턴이 매칭된 카테고리로 문서를 분류한다.

4.2 키워드 집합

키워드 집합(Keyword Set)에 의한 방법은 구문 패턴에 의한 방법과 같이 패턴 매칭을 사용하는 방법으로 패턴 매칭에 사용할 문장 형태를 단어들의 집합으로 표현한다. 패턴의 정의에 사용되는 어휘들을 T_i라 하면 키워드 집합에 의한 분류 규칙의 형태는 다음과 같다.

$$\text{Keyword Set Rule} : \{T_1, T_2, \cdots T_n\} \rightarrow C_k$$

이 방법은 한국어 문장이 많은 경우에 자유로운 어순이 가능하다는 점과, 특정 구문 표현만 분류의 단서가 되는 것이 아니라 특정 단어들이 한 문장 안에 동시에 나타나기만 하면 단어들이 멀리 떨어져 있는 경우에도 단서가 될 수 있다는 점을 고려한 것이다.

키워드 집합 방법에서 사용하는 패턴 매칭 방법의 원리는 유사한 단어로 구성된 문장을 공유하는 문서는 같은 카테고리로 분류될 수 있다는 것이다. 이 때 문장이 문서의 내용을 대표할 수 있는 문장일수록 더욱 같은 카테고리로 분류될 확률이 높다. 예를 들어, 아래의 두 문장은 모두 키워드 집합 {과속, 중앙선, 충돌}에 매칭되며 이러한 문장을 포함한 문서는 같은 카테고리로 분류될 수 있다

문장 1 "빗길을 과속으로 달리던 승용차가 중앙선을 넘어 앞에 오던 화물차와 정면 충돌하였다."

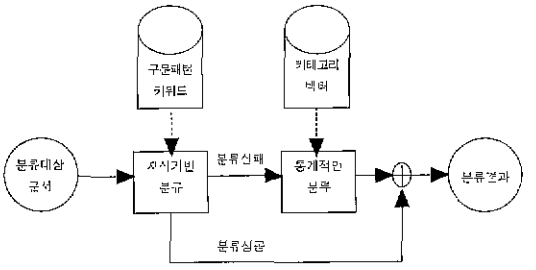
문장 2 : "어제 밤에 일어난 고속도로 정면 충돌 사건의 원인은 과속 주행하던 승용차의 중앙선 침범 때문인 것으로 밝혀졌다."

구문 패턴에 있어서와 마찬가지로 키워드 집합에 있어서도 집합내의 단어 수는 분류의 정확도 및 분류율에 영향을 미친다. 집합내의 단어 수가 많아질수록 그 키워드 집합과 매칭이 되는 문장 수는 적어져서 분류율은 감소하며, 더 엄격한 매칭으로 인해 분류의 정확도는 증가하게 된다.

어떤 문서가 두 개 이상의 서로 다른 카테고리를 나타내는 키워드 집합에 매칭되는 경우에는 매칭된 키워드 집합간에 중요도를 고려하여 더 높은 중요도를 갖는 키워드 집합의 카테고리로 문서를 분류할 수 있다. 키워드 집합간에 중요도를 계산하는 방법에는 단어 집합이 가지고 있는 단어의 개수를 기준으로 하는 방법, 문서의 앞부분에 문서의 내용을 대표하는 문장이 서술될 확률이 높은 점을 고려하여 문장의 앞부분에 매칭된 키워드 집합이 속한 분야에 우선 순위를 주는 방법 등이 있으며, 본 논문에서는 매칭되는 키워드 집합 중 단어의 수가 많은 키워드 집합으로 해당 문서를 분류하는 방법을 사용하였다. 단어의 개수가 많은 키워드 집합일수록 중요하다고 볼 수 있는 이유는 키워드 집합이 가지고 있는 단어가 개수가 많을수록 특정 카테고리를 더 정확하게 표현한다고 할 수 있으며 동시에 그 키워드 집합에 의한 매칭이 오분류일 확률이 상대적으로 낮기 때문이다.

5. 통계적 방법과 지식기반 방법의 통합

2장에서 언급한 바와 같이 통계적 방법과 지식 기반 방법은 상호 보완적인 특징을 가지고 있으므로 두 방법을 적절히 통합하면 분류 성능을 향상시킬 수 있다. 두 방법을 통합하는 방법은 여러 가지가 있을 수 있는데, 통합은 두 방법의 장점을 살리며 단점을 보완하는 방향이 되어야 한다. 즉, 통계적 방법의 장점인 높은 분류율과 지식 기반 방법의 장점인 높은 정확도를 살려서 통계적 방법의 단점인 낮은 정확도와 지식 기반 방법의 단점인 낮은 분류율을 상호 보완 할 수 있도록 통합하여야 한다. 통계적인 문서 분류 방법과 비교해 볼 때, 패턴에 의한 분류 방법은 어느 정도의 의미 분석이 가능하므로, 패턴에 매칭된 문서들의 정확도는 비교적 매우 높으나, 패턴에 매치되지 않는 문서들은 분류되지 않으므로 분류율은 비교적 낮다. 따라서 분류되지 않은 새로운 문서를 분류 할 때, 패턴에 의한 분류 방법으로 1차 분류를 시도한 후, 패턴에 의한 분류 방법으로 분류되지 않은 문서에 대해 통계적인 분류를 시도하면 모든 문서를 분류해 내면서 전체적인 분류 정확도를 높일 수 있다. 본 논문에서는 다음 그림 3과 같이 시스템을 구성하여 분류의 정확도를 높이는 방법을 시도하였다



(그림 3) 분류 시스템의 구성

(그림 3)에서와 같이 전체 분류 시스템은 먼저 분류 대상 문서를 지식 기반 방법으로 분류 한 후, 분류되지 않은 나머지 문서를 통계적 방법으로 분류한다. 이 방법은 문서를 의미 분석 방법과 통계적 방법으로 분류 할 때 두 가지 방법이 한 문서를 서로 다른 범주로 분류하면 의미 분석 방법에 우선을 두어서 분류하겠다는 것이다. 이와 같이 지식 기반 방법에 우선을 두는 것은 지식 기반 방법이 통계적 방법보다 일반적으로 높은 정확도를 나타내므로 타당하다고 할 수 있다. 지식 기반 방법으로 분류되지 않은 문서는 통계적 방법으로 분류되므로 전체적인 시스템은 모든 문서를 분류해 내게 된다.

6. 실험 및 결과

6.1 실험 방법

본 논문에서 제안한 분류 방법들에 대한 성능을 알아보기 위하여 3장에서 제안한 ICF 가중치와 벡터 유사도를 이용한 통계적인 분류 방법에 대한 실험, 구문 패턴과 키워드 집합을 이용한 지식 기반 분류 방법에 대한 실험, 그리고 통계적인 방법과 지식 기반 방법을 통합한 분류 방법에 대한 실험을 수행하였다

실험 대상 문서 집단으로는 94년 조선일보 CD-ROM으로부터 수집한 962개의 경제기사를 이용하였다. 전체 기사는 5개의 대분류 카테고리(level 1)로 분류되며, 이들 각 카테고리에 밑에 총 26개의 소분류 카테고리(level 2)가 있다. 실험집단의 분류는 수작업으로 하였다. 문서의

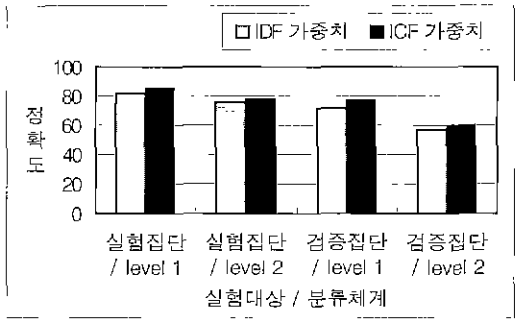
크기는 약 50~200 단어이다. 총 962개의 경제기사 중 723개(75%)를 실험 집단(training set)으로, 나머지 239개(25%)를 검증 집단(test set)으로 하여 실험을 수행하였다.

ICF에 의한 통계적인 분류 방법에 대하여는 계층적 분류체계에 대한 실험을 위해 KTSET에 있는 1000개의 문서를 사용한 실험도 수행하였다. KTSET은 한국어 정보검색 연구를 위해 만들어진 실험용 문서집단으로, 내용은 논문의 요약이고, 제목, 저자, 분류 항목 등이 명시된 양식화된 문서들이다[22] KTSET의 경우는 level 1에 10개, level 2에 57개, level 3에 201개, level 4에 382개의 카테고리가 있는 계층적인 카테고리 구조로 분류되며, 실험집단의 분류는 각 문서에 이미 명시된 카테고리들을 이용하였다. 문서의 크기는 요약 부분이 약 100 단어 정도이다. KTSET에서는 제목과 요약 부분만을 실험대상 문서로 하였다. 총 1000개의 문서 중 750개(75%)를 실험집단으로, 나머지 250(25%)개를 검증집단으로 하였다.

문서에서의 색인어 추출에는 한성대학교에서 개발한 형태소 분석기 HAM[21]을 사용하였다

6.2 통계적인 분류 실험

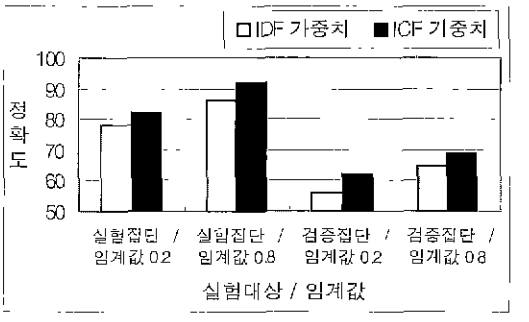
통계적인 분류에 대한 실험은 평면적인 구조를 갖는 분류체계와 계층적 분류체계 두 경우를 대상으로 하여 분류 실험을 수행하여 본 논문에서 제안한 ICF 가중치를 사용한 경우와 IDF 가중치를 사용한 경우의 분류 정확도를 비교하였다.



(그림 4) 통계적인 분류의 정확도 - 경제기사 (평면적 분류체계)

경제기사를 대상으로한 평면적인 분류체계에서의 분류실험 결과를 (그림 4)에 나타내었다. 각 실험결과는 실험집단과 검증집단으로 나누어 측정하였고, 분류체계의 각 level을 하나의 평면적인 분류체계로 보고 각 level에 대하여 독립적으로 분류 실험을 수행하였다.

실험집단의 결과를 보면 실험집단과 검증집단 모두 ICF를 사용한 경우가 level 1에서 약 4%, level 2에서 약 3% 높은 정확도를 보이고 있다. 하위 level에서 정확도 향상이 비교적 낮은 이유는 하위 level일수록 카테고리 수는 증가하는 반면 각 카테고리에 속한 문서 수는 줄어들어 ICF 계산을 위한 데이터가 적기 때문이다. 이 결과를 볼 때 단순한 평면적 분류체계에서도 문서간의 분리도를 나타내는 IDF보다 카테고리간의 분리도를 나타내는 ICF의 사용이 분류의 정확도를 향상시킨다는 것을 알 수 있다.



(그림 5) 통계적인 분류의 정확도- KTSET (계층적 분류 체계)

KTSET을 대상으로한 계층적인 분류체계에서의 분류실험 결과는 (그림 5)에 나타내었다 이 실험에서는 3장에서 설명한 분류 알고리즘에 따라 분류를 수행하였다. 즉, 각 level에서의 분류 시 상위 분류 결과의 부속카테고리들에 대해서만 그 범위를 한정시켜 ICF를 동적으로 계산하였고, 유사도가 일정 임계값을 넘지 못하면 분류를 종료하도록 하였다. 정확도 계산은 문서가 복수의 항목으로 분류되어 있는 경우 실험 결과가 이들 항목 중에 포함되어 있으면 맞는 결과로 하였고, 또한 실험 결과가 분류 항목과 정확히 같거나 분류 항목의 상위 카테고리이면 맞는 결과로, 하위 카테고리이면 틀린 결과로 간주하였다.

(그림 5)의 결과를 보면 임계값을 높게 할수록 좋은 결과를 보이는데, 이는 임계값이 높으면 보다 높은 level에서 분류가 중단되므로 앞에서 설명한 계층적 분류체계에서의 정확도 계산 방법상 올바른 분류가 될 확률이 높기 때문이다

계층적 분류 실험에서도 실험집단과 검증집단 모두 ICF를 가중치로 이용한 경우가 좋은 결과를 보였으며, 특히 계층적 분류 실험의 경우, 3%~4%의 정확도 향상을 보인 평면적 분류체계에서의 분류 결과보다 높은 4%~6%의 정확도 향상을 보임으로서 ICF 가중치의 사용이 계층적 분류체계에서의 분류에 더욱 효과적임을 알 수 있다.

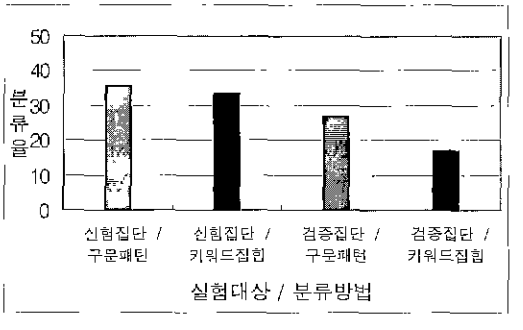
6.3 지식 기반 분류 실험

지식 기반 분류 방법에 대한 실험은 경제기사를 대상으로 구문 패턴에 의한 분류 실험과 키워드 집합에 의한 분류 실험을 수행하였다. 실험에 사용한 구문 패턴은 총 171개 였으며, 키워드 집합은 한 단어로 구성된 집합이 110개, 두 단어로 구성된 집합이 269개, 세 단어 이상으로 구성된 집합이 191개 였고, 각 실험결과는 실험집단과 검증집단으로 나누어 측정하였다. 지식 기반 방법은 매칭이 안되는 문서들은 분류할 수 없으므로 분류에 성공한 문서의 비율인 분류율을 정확도와 함께 나타내었다. 본 실험에서 측정한 분류율과 정확도의 정의는 식(4) 및 식(5)와 같다.

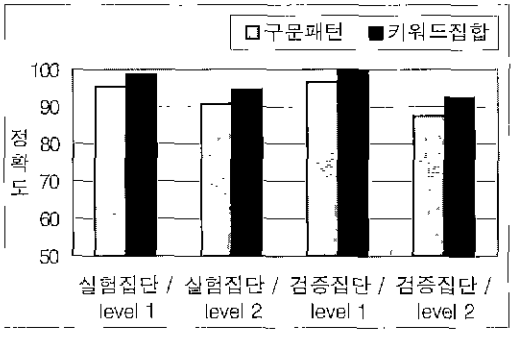
$$\text{정확도} = \frac{\text{바르게 분류된 문서 수}}{\text{분류된 총 문서 수}} \tag{4}$$

$$\text{분류율} = \frac{\text{분류된 총 문서 수}}{\text{총 문서 수}} \tag{5}$$

지식 기반 분류 실험 결과를 (그림 6)과 (그림 7)에 나타내었다. 실험 결과를 보면 검증집단의 경우 패턴을 이용한 지식 기반 분류는 두 가지 방법 모두 분류율은 30% 이하로 낮으나, 정확도는 80% 이상으로 매우 높은 것으로 나타났다.



(그림 6) 지식 기반 분류의 분류율



(그림 7) 지식 기반 분류의 정확도

6.4 복합적인 방법에 의한 분류 실험

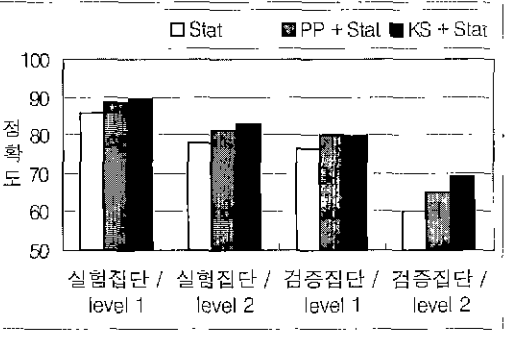
복합적인 분류 방법에 대한 실험은 같은 문서집단을 대상으로 3장에서 제안한 방법에 의한 분류 실험을 하여 통계적인 분류만을 수행한 결과와 정확도를 비교하였다. 본 실험에서 측정한 복합적인 방법의 전체 정확도 P_{total}은 패턴에 의한 분류의 정확도를 P_p, 분류율을 C_p, 통계적 방법에 의한 분류의 정확도를 P_s라고 하면 식(6)과 같다.

$$P_{total} = C_p \cdot P_p + (1 - C_p) \cdot P_s \tag{6}$$

〈표 2〉 복합적인 방법에 의한 분류 실험 결과

분류 방법	실험집단				검증집단			
	level 1		level 2		level 1		level 2	
	정확도	정확도 증가율	정확도	정확도 증가율	정확도	정확도 증가율	정확도	정확도 증가율
Stat	85.6%	-	78.1%	-	76.5%	-	60.3%	-
PP + Stat	88.5%	3.4%	81.4%	4.2%	79.9%	4.4%	65.1%	8.0%
KS + Stat	89.8%	4.6%	83.0%	6.3%	79.5%	3.9%	69.2%	14.8%

<표 2>에 통계적인 방법에 의한 분류 실험 결과(Stat)와 통계적인 방법에 본 논문에서 제시한 패턴에 의한 분류 방법을 접목시킨 복합적인 방법에 의한 실험 결과(PP+Stat, KS+Stat)를 나타내었다. 실험 결과를 보면 본 논문에서 제안한 복합적인 방법을 사용했을 경우, 구문 패턴을 이용한 방법은 통계적인 방법만을 사용한 경우에 비해 실험집단에서 3~4%, 검증집단에서 4~8% 정도 정확도가 향상되었고, 키워드 집합을 이용한 방법은 통계적인 방법만을 사용한 경우에 비해 실험집단에서 5~6%, 검증집단에서 4~15% 정도 정확도가 향상되었다. 특히 복합적인 분류 방법은 카테고리가 세분화되어 있는 경우(level 2)에 더 높은 정확도 향상을 보였다. 그림 8에 종합적인 결과를 그래프로 나타내었다.



(그림 8) 복합적인 방법에 의한 분류의 정확도

7. 결 론

본 논문에서는 통계적인 문서 분류에 문장의 의미를 파악할 수 있는 패턴들을 이용한 지식 기반 분류 방법을 접목시킴으로써 분류의 정확도를 높이는 방법을 제안하였다. 통계적인 분류 방법으로는 벡터 유사도에 의한 분류 방법을 사용하였으며, 단어에 대한 새로운 가중치 계산 방법으로 역카테고리빈도(ICF)를 정의하고 이를 이용한 계층적 분류체계에서의 분류 방법을 제안하였다. 지식 기반 분류 방법으로는 문장 형태를 표현하는 방법에 따라 구문 패턴에 의한 방법과 키워드 집합에 의한 방법을 제안하였다.

통계적인 분류에 대하여는 조선일보 경제기사와 KTSET을 대상으로한 분류 실험을 통하여 ICF를 사용한 경우가 IDF를 사용한 경우보다 평면적인 분류체계와 계층적인 분류체계에서 모두 더 정확한 분류를 한다는 것을 보였고, 지식 기반 분류에 대하여는 조선일보 경제기사를 대상으로한 분류 실험을 통하여 지식 기반 방법이 분류율이 낮은 반면 정확도는 매우 높음을 보였으며, 통계적인 분류와 지식 기반 분류를 복합적으로 사용한 분류 실험을 통하여 통계적인 방법만을 사용한 경우에 비해 본 논문에서 제안한 분류 방법을 사용한 경우가 높은 정확도를 나타냄을 보였다.

향후 과제로 지식 기반 방법의 확장성을 높이기 위해 패턴을 찾아내는 작업을 자동화하는 방법에 대한 연구가 필요하며, 단순한 텍스트 문서뿐 아니라 HTML과 같은 구조화된 문서의 분류 방법에 대한 연구도 수행되어야 할 것이다.

참 고 문 헌

[1] M Blosseville. G. Hebrail, M Monteil, and N Penot., "Automatic document classification : natural language processing, statistical analysis, and expert system techniques used together," *SIGIR'92*, 1992.

[2] W. Frakes. and R. Baeza-Yates, *Information Retrieval*, Prentice Hall, 1992.

[3] N. Fuhr. "Models for retrieval with probabilistic indexing," *Information Processing and Management*, 25(1), 1989.

[4] K. Hamill and A Zamora. "The Use of Titles for Automatic Document Classification," *Journal of the American Society for Information Science*, 1980

[5] D. Harman, "Ranking algorithms," in *Information Retrieval Data Structures and Algorithms*, Prentice Hall, 1992.

[6] P. Hayes and S. Weinstein, "CONSTRUE/TIS : A system for content-based indexing of a database of news stories," *Second Annual Conference on Innovative Applications of Artificial Intelligence*, 1990

[7] P. Hayes, P. Anderson, I. Nirenburg, and L Schmandt "TCS : A Shell for Content-based Text Categorization," *Proceedings of the 6th IEEE Conference on Artificial Intelligence Applications*, Santa Monica, March 1990.

[8] J. R Hobbs., D Appelt, M. Tyson, J. Bear and D. Israel, "FASTUS : System summary," *Proceedings of Fourth Message Understanding Conference*, 1992

[9] R Hoch., "Using IR techniques for text classification in document analysis," *SIGIR'94*, 1994.

[10] P. Jacobs., *Text-Based Intelligent Systems*, Lawrence Erlbaum, 1992

[11] P Jacobs., "Using statistical methods to improve knowledge-based news categorization," *IEEE Expert*, April, 1993.

[12] J. Hobbs, D. Appelt, J. Bear, D. Israel, and M. Tyson "FASTUS : A System for Extracting Information from Natural-Language Text"

[13] L Larkey and W. Croft, "Combining classifiers in text categorization," *SIGIR'96*, 1996.

[14] D. Lewis. "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," *SIGIR'92*

[15] D. Lewis. "Evaluation and optimizing autonomous text classification system," *SIGIR'95.*

[16] D. Lewis., R. Schapire, and J. Callan, "Training algorithms for linear text classifiers," *SIGIR'96.*

[17] B Masand, "Classifying News Stories using Memory Based Reasoning," *SIGIR'92*

[18] M. Maron, "Automatic indexing : An experimental inquiry." *Journal of the ACM*, 1961.

[19] Proceedings of the Fourth Message Understanding Conference. Morgan Kaufmann. CA 1992.

[20] G. Salton *Automatic Text Processing : The Trans-*

formation, Analysis, and Retrieval of information by Computer. Addison Wesley Publishing Co., 1989.
[21] 강승식, 이하규, "한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능", 한글 및 한국어 정보처리 학술 발표논문집, 1996.
[22] 김재군, 김영환, 김성혁, "한국어 정보검색 연구를 위한 시험용 데이터 모음 KTSET 개발", 한글 및 한국어 정보처리 학술 발표논문집, 1996
[23] 박미경, 김민정 "부분 파싱을 이용한 한국어 명사구, 술어구와 접사의 색인 기법", 정보과학회 학술발표 논문집, 4, 1997
[24] 송재관, 홍성용, 박찬곤 "기계 번역을 위한 한국어 문장 패턴에 관한 연구", *정보과학회 학술발표 논문집*, 10. 1996.
[25] 엄미현, 신대규, 나동렬 "한국어의 구조적 애매성", 정보과학회 학술발표 논문집, 4. 1996.
[26] 임해창, 임희석, 윤보현, "자연어처리 연구 동향 : 통계 기반의 자연어 처리", 한국정보과학회지, Vol.12, No.9, pp.20-30, 1994
[27] 정영미, 정보검색론, 구미무역 출판부, 1993.
[28] 조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류", *정보과학회 학술발표 논문집*, 4, 1997.
[29] 최동시, 정경택, "카테고리와 키워드의 밀접성 정보에 의한 문서 자동 분류 시스템 설계 및 구현", 정보과학회 학술발표 논문집, 10, 1995.

한 정 기
e-mail : jhan@dgu.ac.kr
1997년 동국대학교 컴퓨터공학과 졸업(학사)
1999년 동국대학교 대학원 컴퓨터공학과 졸업(석사)
1999년~현재 동국대학교 대학원 컴퓨터공학과 박사과정 재학
관심분야 : 기계학습, 웹 에이전트

박 민 규
e-mail : mpark@dgu.ac.kr
1997년 동국대학교 컴퓨터공학과 졸업(학사)
1999년 동국대학교 대학원 컴퓨터공학과 졸업(석사)
1999년~현재 웹 패턴 테크놀로지 개발실 근무
관심분야 : 지능형 검색엔진, HCI, 웹 에이전트

조 광 제
e-mail : gjcho@dgu.ac.kr
1996년 호서대학교 컴퓨터공학과 졸업
1998년 동국대학교 대학원 컴퓨터공학과 졸업(석사)
1998년 현재 서울시스템 DTP사업본부 개발팀 근무
관심분야 : 자연언어처리, 정보검색, 자동문서분류

김 준 태
e-mail : jkim@dgu.ac.kr
1982년 서울대학교 제어계측공학과 졸업(학사)
1988년 미국 University of Southern California 전기공학전공(M.S.)
1993년 미국 University of Southern California 컴퓨터공학전공(Ph.D.)
1994년~1995년 미국 Southern Methodist University Postdoc
1995년~현재 동국대학교 컴퓨터공학과 조교수
관심분야 : 인공지능, 정보검색, HCI, 데이터마이닝